DNA Microarray Data Classification via Haralick's Parameters

Kalpana Ramakrishnan[1], Perambur S. Neelakanta*[2]

[1]Department of Biomedical Engineering,
Rajalakshmi Engineering College, Chennai, India.
[2]Computer and Electrical Engineering Department, Florida Atlantic University
Boca Raton, Florida 33431, USA
[1]kalpanatirth@gmail.com; *[2]neelakan@fau.edu

Abstract

Addressed in this paper is an information-theoretics inspired entropy co-occurrence approach to extract subset features denoting relative extents of gene expression levels in a DNA microarray data. Experimentally deduced microarray data (via wet-lab studies) exhibits intensity gradation seen as yellow, green and red spots. Such distinguishable (color) features of microarray patterns depict the associated heterogeneity of meso textures and their spatial distribution manifested as intensity gradation in the colored spots. Normally, a procedure is sought to distinguish two such patterns of microarrays in which the colored spots are distinctly expressed. For this purpose, considered in this study is the so-called Haralick's feature-finding algorithms and deducing Haralick's feature scores in terms of the entropy co-occurrence information in an image with reference to a threshold profile. The scores obtained from intensity gradations of colored spots can be correlated to the relative extent of the color status of the spots. Hence corresponding gene expression (depicting tumor, non-tumor or mixed condition etc. as implied by proportional abundance of color pixels in the microarray) is quantitatively indicated. Relevant strategy in the use of Haralick's algorithms in microarray contexts is a new and novel approach. Details on simulation experiments using synthetic microarray patterns and results obtained thereof are presented to illustrate the efficacy of the proposed method.

Keywords

Dna Microarrays; Pattern Classification; Haralick'S Parameters

Introduction

The living system consists of innumerable cells containing the alphabets of life inscribed as genes. The translation of information encoded in a gene is termed as gene expression and this information is eventually used in making protein (or transcribed into specific RNAs like transfer and ribosomal RNAs needed for certain operations in the cell). In pursuant of the steps of central dogma of microbiology, the expressed genes depicting DNA segments are transcribed into messenger RNA (mRNA), (which is then translated into protein, or transcribed into transfer and ribosomal RNAs).

A small analytical device (termed as the "microarray"). functions like a biological microprocessor allowing genomic exploration *via* rapid quantitative analysis of gene expression patterns and other related entities. It provides a tool to potentially identify and quantitate levels of gene expression for all genes in an organism. Microarray analysis implies a blend of computational and experimental procedures adopted to explore and extract gene-related biological, chemical and physical features. With the advent of explorative oncology adopted to classify different versions of cancer *vis-à-vis* the patterns of gene activity in tumor cells, the significance of DNA microarray designs has been widely emphasized (Schena, 2001; Shapiro and Stockman, 2001; Stekel, 2003; Ewis, et al., 2005; Nair, 2008).

The basics of microarray technology are illustrated in Figure 1. In essence, it involves the realization of a microarray of small spots of DNA (arranged as an ordered array) fixed to a glass-slide or a nylon membrane. It deliberates a cDNA hybridization experiment leading to comparing the amounts of different RNAs in two cell populations. The associated wet-lab efforts involve first making of a planar substrate of a glass or a silicon chip (using dry chemistry and automation), which is decorated with a rectangular array of thousands of circular spot, each of which a unique nucleotide sequence of complementary DNA (cDNA) corresponding to the test mRNA (Baggerly 2001).

The microarray realized via wet-lab technique can be used to assay gene expression within a single sample or to compare gene expression in two different cell types or tissue samples, such as in healthy and diseased tissue. Suppose there are two cells: cell-type 1, a healthy cell, and cell-type 2, a diseased cell, both of which contain an identical set of four genes, called A, B, C, and D. It is of interest to determine the expression profile of these four genes in the two cell types. Thus, the mRNA is isolated from each cell type and used as templates to generate cDNA with a "fluorescent tag" attached. That is, two-color microarrays (or two-channel microarrays) are hybridized with cDNA prepared from the two samples being compared (namely, healthy cell type 1 and diseased cell type 2); and they are labeled distinctly with two different fluorophores.

Fluorescent dyes commonly used for cDNA labeling include Cy3, which has a fluorescence emission wavelength of 570 nm (corresponding to the green part of the light spectrum), and Cy5 with a fluorescence emission wavelength of 670 nm (corresponding to the red part of the light spectrum). The two Cy-labeled cDNA samples are mixed and hybridized to a single microarray and then scanned in a microarray scanner to visualize fluorescence of the two fluorophores (after excitation with a laser beam of a defined wavelength).

In the two-colored labeled samples, the green color represents the control DNA (where either DNA or cDNA derived from normal tissue is hybridized to the target DNA). The red color denotes the sample DNA, (where either DNA or cDNA is derived from diseased tissue hybridized to the target DNA); and the yellow color represents a combination of control and sample DNA (where both are hybridized equally to the target DNA). Further, black depicts areas where neither the control nor sample DNA is hybridized to the target DNA. Thus, each spot on an array is associated with a particular gene. Each color in an array represents either healthy (control) or diseased (sample) tissue. Depending on the type of array used, the location and intensity of a color will tell us whether the gene or mutation is present in either the control and/or sample DNA, which will also provide an estimate of the expression level of the gene(s) in the sample and control DNA.

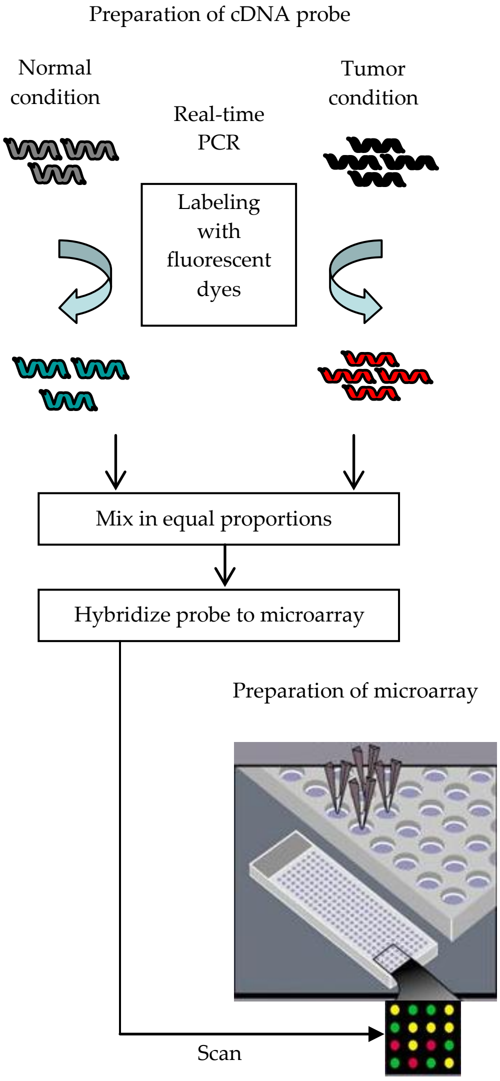

Preparation of cDNA probe
Normal condition
Real-time PCR
Tumor condition
Labeling with fluorescent dyes
Mix in equal proportions
Hybridize probe to microarray
Preparation of microarray
Scan

FIG. 1 DNA MICROARRAY TECHNOLOGY

A scanner is used to measure the pattern of variations in the fluorescence of the sample on the slide and prepares the soft-copy of the microarray database of the chip. Hence, quantitative gene expression data can be obtained by determining the fluorescence intensity at each colored microarray location. Figure 2 illustrates a typical microarray showing varying degrees of fluorescence intensity. These intensities can then be coded to a color palette (along a horizontal bar).

By using microarray patterns (similar to Figure 2), they can be analyzed to elucidate the comparative features of normal *versus* tumor genes. Objectively, such analysis is done to quantify the expression of large arrays of genes at discrete intervals of time.

The end-product of the comparative hybridization experiment in microarrays described above refers to an image-array and the underlying data is obtained by scanning the array.

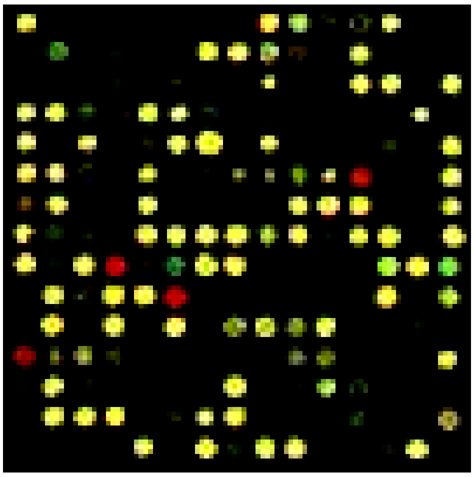


FIG. 2 ILLUSTRATION OF A MICROARRAY

The present study proposes a relevant method of analysis towards assaying the scanned microarray patterns specified as images in digital format. Pertinent details are as follows:

DNA Microarray Data Classification Via Haralick'S Coefficients

A DNA microarray pattern captured via laboratory techniques described above when denoted as a digital image would correspond to the pixel values represented by binary digits 0 and 1. These binary values can be represented by a matrix array of rows and columns of cells called, pixels. Further, the DNA microarray is a trispectral image containing pixels of red, green and yellow colors. Relevant 2D-pixel image M[x, y], has a vector of value at each spatial point or pixel. Due to the color set being {red, green, yellow}, such a vector has three disctinct representations. Therefore, given a test-image, it can be compared and contrasted against another (or a template) in terms of their trispectral constituent vectors and the spectral entities implicitly depict the characteristic image-features associated with the test images.

Appropriate algorithms and computational methods can be prescribed to extract the underlying feature details in the mixed state of colored pixels of DNA microarray images; and, in the context of the present study, interpreting such DNA microarray data that exhibits intensity gradation seen across the trispectral yellow, green and red spots is attempted via the the associated entropy co-occurrence details. That is, a procedure is sought to distinguish two pixel patterns in which the colored spots are distinctly expressed; and, relevant gene expression details correspond to information borne by the entropy characteristics of the co-occurrence of the pixels in the layout of the microarray. The list of relevant steps pursued is as follows:

- Firstly, the trispectral image is read and corresponding grey-level image is generated. Then the grey-level image generated is binary formatted in specific categories *via* a "thresholding" operation, in which some of the pixels are selected as the foreground entities that make up the objects of interest and the rest as background pixels. That is, for the prevailing grey-tone gradation in the image, certain (grey-tone) values are prescribed as threshold values so as to separate the pixels into groups.
- Separating (classifying) the pixels in a DNA microarray into groups: This procedure involves thresholding described above. Suppose the *extrema* of grey-scale namely, white and black conform to values 0 and 1 respectively. Then the yellow color (closer the tinge of white) can be classified as the category corresponding to a grey-level threshold of 0.1 (closer to 0). Likewise, the red color (closer the tinge of black) can be classified as the category corresponding to a grey-level threshold of 0.9 (closer to 1). The green color signifies a transition from yellow-to-red with grey-level fixation at 0.5.

As mentioned earlier, the texture details of multi-spectral bio-images like (DNA microarrays) implicitly depict different grades of pathology (like cancer malignancy) (Bala, 2011, Bryant et al., 2004; Ward, 2006); and the associated grey-levels probabilistically estimate the features of the artifacts in the image. That is, the grey-level features dubbed from the trispectral DNA microarray and presented in binary format denote the intensity properties of the image texture pixel-by-pixel; and the spatial occurrence of the grey-levels (being stochastic in nature) would pose correspondingly a probabilistic attribute to the grey-level observed at each pixel.

The pseudocode (# 1) is indicated on the DNA microarray data preprocessing.

```
--------------------------------------------------------------------------------
Pseudocode # 1:   Microarray data preprocessing and binary
                  formatting the image
Initialize
    → Colored scanned image is read using the command
      'imread' in Matlab™
        → Corresponding map of grey level image is generated
          and stored
%% In terms of grey-scale, all-black-pixels correspond to scaling at
   grey-level '1'. On the other extreme side, when all the pixels are
   white, it is graded as '0'. In between, the yellow-rich image
   would correspond to a grey-level closer to zero, say, 0.1; and,
   the red-rich image would be closer to 1, say 0.9; lastly, the
   green-rich condition can be taken as a grey-level in the vicinity
   of 0.5.
        → The grey level image is then binary formatted in
          following categories:
        ← Zero threshold category: $t_T = 0$
%%  Zero-level threshold category implies fluorescence depicting
    white color
        ← First category corresponds to a grey-level threshold:
          0.1
%%  First category implies fluorescence depicting the high-
    intensity yellow color
        ← Second category corresponds to a grey-level
          threshold: 0.25
%%  Second category implies fluorescence depicting the low-
    intensity yellow color on transition to green
        ← Third category corresponds to a grey-level threshold:
          0.50
%%  Third category implies fluorescence depicting the high-
    intensity green color
        ← Fourth category corresponds to a grey-level
          threshold: 0.75
%%  Fourth category implies fluorescence depicting the low-
    intensity green color on transition to red
        ← Fifth category corresponds to a grey-level threshold:
          0.90
%%  Fifth category implies fluorescence depicting the high-
    intensity red color
        ← Maximum threshold category: $t_T = 1$
%%  Maximum-level threshold category implies fluorescence
    depicting black color
Store
    → Store as arrays of binary formatted classified set of grey
      level images
End
--------------------------------------------------------------------------------
```

With the grey-level based categorized images specified in binary format as above, each can then be subjected to a scoring procedure wherein the scores obtained are interpreted as the relative extent of the population of yellow, green, red pixels in the microarray. For scoring purpose, the so-called Haralick's feature-finding algorithms are considered in this study as outlined below:

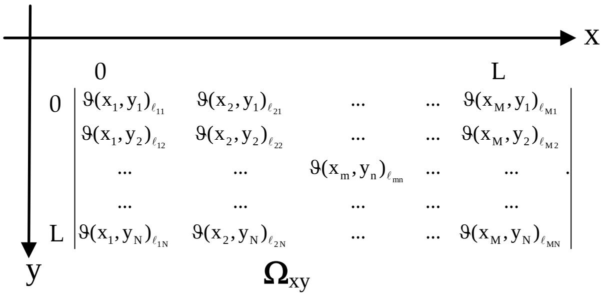
Haralick et al. (1973) proposed fourteen variants of statistical coefficients that could be deduced from a matrix ascribed to the texture map of an image. That is, Haralick's parameters can discriminate image features in terms of the associated textural characteristics of images taken in a matrix format. Relevant considerations are popular in medical image contexts, for example, as indicated in (Kalpana and Muttan, 2012; Kalpana et al., 2011; Chaddad et al., 2011). However, rare known prior studies are available in open-literature (to the best of the authors' knowledge) in adoption of Haralick's coefficients in the exclusive task of characterizing image textures of DNA microarrays (as conceived in the present work). As such, the strategy of invoking Haralick's algorithm in the context of DNA microarray analysis is new and novel.

To compute Haralick's coefficients, an algorithm known as the grey-level co-occurrence matrix (GLCM) which is first constructed consistent with the set of pixel data of the digital 2D-pattern in question refers to a tabulation of how often different combinations of grey levels simultaneously occur in the image or a part of it. The underlying classification of different grades can then be extracted *via* one of the fourteen variants of statistical coefficient measures due to Haralick indicated in (Haralick et al., 1973). Pursued here is the entropy parameter coefficient measure which will be described later. The computational approach of GLCM construction is as follows:

To construct the GLCM of an image, a displacement vector **d** with reference to a pixel (expressed by a radial distance δ and an orientation angle (θ) is first defined so as to decide on the relative disposition of pixel-to-pixel locations and the extent of prevailing grey-level correlations in the test image-array.

Concerning the choice of the orientation angle θ, every pixel can have eight neighboring pixels allowing θ to be: (0°, 45°, 90°, 135°, 180°, 225°, 270° or 315°). However, considering the diagonal symmetry of matrix format of GLCM, the co-occurring pairs obtained by choosing θ equal to 0° is same as θ being 180°. Likewise, this symmetry concept extends to 45°, 90° and 135° as well. Therefore, the choice on θ is limited to four. Further, the dimension of a GLCM is determined by the maximum grey value of the pixel and the number of grey-levels adopted. It is crucial to determine the efficacy of GLCM computation. More levels would mean more accurate depiction of textural information, but it also implies more computational burden.

Figure 3 illustrates the 2D-plane of an image-space Ω_{xy}. Relevant pixel-by-pixel grey-level intensity (measured) across this 2D-profile is denoted by $\vartheta(x_m, y_n)$ at any coordinate (x_m, y_n) and the scale of intensity level is indicated by ℓ_{mn}. Further, the size of GLCM is taken as (M × N) with m = 1, 2, …, M; and n = 1, 2, …, N.



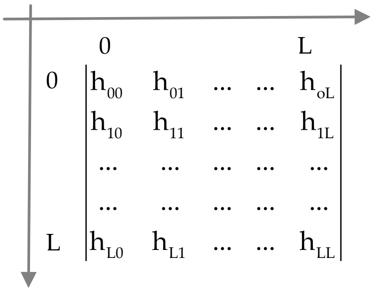
$$\begin{array}{c|ccccc|} & 0 & & & & L \\ 0 & \vartheta(x_1,y_1)_{\ell_{11}} & \vartheta(x_2,y_1)_{\ell_{21}} & \cdots & \cdots & \vartheta(x_M,y_1)_{\ell_{M1}} \\ & \vartheta(x_1,y_2)_{\ell_{12}} & \vartheta(x_2,y_2)_{\ell_{22}} & \cdots & \cdots & \vartheta(x_M,y_2)_{\ell_{M2}} \\ & \cdots & \cdots & \vartheta(x_m,y_n)_{\ell_{mn}} & \cdots & \cdots \\ & \cdots & \cdots & \cdots & \cdots & \cdots \\ L & \vartheta(x_1,y_N)_{\ell_{1N}} & \vartheta(x_2,y_N)_{\ell_{2N}} & \cdots & \cdots & \vartheta(x_M,y_N)_{\ell_{MN}} \end{array}$$

FIG. 3 A GLCM DEPICTING THE 2D-MAP OF THE INTENSITIES OF GREY-LEVELS ALONG X AND Y DIRECTIONS OF A TEST IMAGE-SPACE, Ω_{XY}

Consider two pixels at the coordinates $(x_a, y_b) \in \{x_m, y_n\}$ and $(x_c, y_d) \in \{x_m, y_n\}$ in the (x, y)-space in the 2D-matrix of Figure 3. Let the measured grey-level intensities at these locations be $\vartheta(x_a, y_b)_\ell$ and $\vartheta(x_c, y_d)_\ell$ respectively. Further, 0 ≤ mn ≤ L denotes the scale of integer values of the grey-level at (x_m, y_n), (with zero representing the white-level and L denotes the maximum grey-level towards black). The Euclidean distance between (x_a, y_b) and (x_c, y_d) is denoted by the vector $\mathbf{d}_{ab\text{-}cd}$. (In *lieu* of the Euclidean distance, other statistical distance measures such as, Mahalanobis distance etc. can also be adopted without any loss of generality. However, the Euclidean distance is chosen here for its simplicity in demonstrating the scope of the study).

By using ϑ-matrix of the (x, y) plane, namely, $[\{\vartheta(x_m, y_n)_\ell\}]$ as illustrated in Figure 3, four other matrices can also be constructed involving the slant-angle as follows:

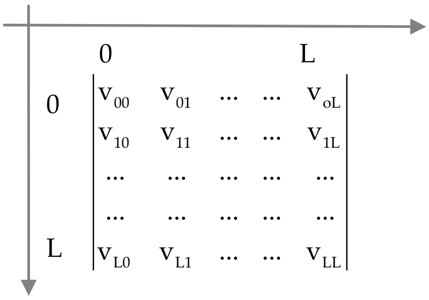
Horizontal (= 0°)-matrix (H_o-matrix): Consider any arbitrary element $\vartheta(x_m, y_n)_\ell$ in the ϑ-matrix and note its grey-scale (integer) value mn lying in the range 0 ≤ ◎ mn ≤ L. Scan the entire row (horizontal scan) containing this element and count the number of times ◎ mn value co-occurring as the neighbor (implying |**d**| =1), 0, 1, …, and L. The count value is denoted as h. This is repeated for each row starting with11 12…, 1N and the resulting H_o-matrix is constructed as shown in Figure 4.



$$\begin{array}{c|ccccc|} & 0 & & & & L \\ 0 & h_{00} & h_{01} & \cdots & \cdots & h_{oL} \\ & h_{10} & h_{11} & \cdots & \cdots & h_{1L} \\ & \cdots & \cdots & \cdots & \cdots & \cdots \\ & \cdots & \cdots & \cdots & \cdots & \cdots \\ L & h_{L0} & h_{L1} & \cdots & \cdots & h_{LL} \end{array}$$

FIG. 4 REPRESENTATION OF: H_O-MATRIX

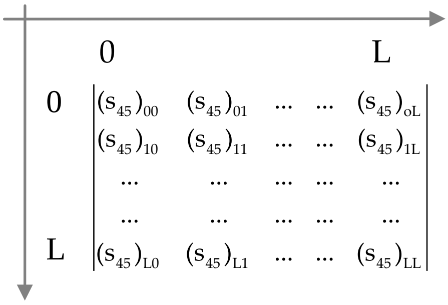
Vertical (= 90°)-matrix (V_{90}-matrix): Similar to H_o-matrix, by scanning vertically along each column of the ϑ-matrix, the resulting V_{90}-matrix is obtained as illustrated in Figure 5, where v stands for the count of co-occurring values along vertical direction and horizontal directions respectively.



$$\begin{array}{c|ccccc|} & 0 & & & & L \\ 0 & v_{00} & v_{01} & \cdots & \cdots & v_{oL} \\ & v_{10} & v_{11} & \cdots & \cdots & v_{1L} \\ & \cdots & \cdots & \cdots & \cdots & \cdots \\ & \cdots & \cdots & \cdots & \cdots & \cdots \\ L & v_{L0} & v_{L1} & \cdots & \cdots & v_{LL} \end{array}$$

FIG. 5 REPRESENTATION OF: V90-MATRIX

Slant (= 45°) and (= 135°)-matrix (s_{45} and s_{135}-matrix): A diagonal count matrix along 45°-slant and 135° slant can also be specified as shown in Figures 6(a) and 6(b), where s_{45} and s_{135} denote the count of co-occurring values diagonally along 45° and 135° respectively with respect to an arbitrary element $\vartheta(x_m, y_n)_\ell$.



$$\begin{array}{c|ccccc|} & 0 & & & & L \\ 0 & (s_{45})_{00} & (s_{45})_{01} & \cdots & \cdots & (s_{45})_{oL} \\ & (s_{45})_{10} & (s_{45})_{11} & \cdots & \cdots & (s_{45})_{1L} \\ & \cdots & \cdots & \cdots & \cdots & \cdots \\ & \cdots & \cdots & \cdots & \cdots & \cdots \\ L & (s_{45})_{L0} & (s_{45})_{L1} & \cdots & \cdots & (s_{45})_{LL} \end{array}$$

FIG. 6(a) REPRESENTATION OF: S_{45}-MATRIX

$$\begin{array}{c|ccccc|}
 & 0 & & & & L \\
0 & (s_{135})_{00} & (s_{135})_{01} & \cdots & \cdots & (s_{135})_{oL} \\
 & (s_{135})_{10} & (s_{135})_{11} & \cdots & \cdots & (s_{135})_{1L} \\
 & \cdots & \cdots & \cdots\ \cdots & & \cdots \\
 & \cdots & \cdots & \cdots\ \cdots & & \cdots \\
L & (s_{135})_{L0} & (s_{135})_{L1} & \cdots & \cdots & (s_{135})_{LL}
\end{array}$$

FIG. 6(b) REPRESENTATION OF: S135-MATRIX

The four matrices indicated above denote implicitly the intensity level count variations in terms of the underlying grey-level gradient; hence, by determining such intensity variations along the four directions will account for the associated anisotropy of the image. Further, an average of the four co-occurrence matrices as above can be taken to ensure rotational invariance as suggested by (Caban, et al., 2007). However, in the present study, each of the four matrices is considered separately in order to evaluate the statistical feature map of the test image under discussion on ensemble basis; and the statistical feature evaluated for the four matrices are then averaged. This would maintain the rotational invariance and the anisotropy considerations without any loss of generality.

For any given category of threshold setting (0 through 1) indicated earlier and the GLCM constructed thereof, the next task is to quantify the textural features in terms of the probabilities estimated. This can be done by resorting to one of the fourteen Haralick's feature parameters defined in (Haralick, et al., 1973) which corresponds to the entropy coefficient (ENT) and is a descriptor of randomness or uncertainty of the grey-level variations in the test pattern. Essentially, entropy concept is based on Shannon's information-theoretic considerations. The underlying details are indicated in the literature (Wang, 2008; Furlanello, et al., 2003; Zhu, et al., 2010; Bala, 2011) for microarray-specific classification of gene expressions. However, relevant pursuits do not follow the GLCM and Haralick's coefficient based approach.

The entropy pararmeter (ENT) of the set of Haralick's coefficients can be expressed in the units (bits) of Shannon's information measure as follows:

$$I = ENT = -\sum_{i=1}^{M}\sum_{j=1}^{N}[P(i,j)]_{r,d} \qquad {}_{2} \qquad {}_{r,d} \qquad (1)$$

where [M × N] denotes the size of the GLCM matrix with M or N depicting the extents of grey-level of the image and in GLCM representation, $[P(i, j)]_{r,d}$ depict the probabilities of transition from pixel of an i^{th} grey intensity to a pixel of a j^{th} intensity separated by a translational vector set defined by: {direction r (expressed in terms of) and distance, **d**}. For computational purpose, the subsets d = {1, 2, 3, 4} and r: = {0°, 45°, 90°, 135°} are considered and hence, the procedure of image characterization using ENT parameter is described via a pseudocode (# II) as indicated in three parts, A, B and C

Peudocode# II: Comparison of microarray data: Pseudocode on constructing the GLCM and evaluating the Haralick's parameter

Part A: **Initialization**

Read

→ Read and retrieve the stored array of binary formatted grey-level test image(s)

Choose

→ Threshold setting: t_T

t_T	Description
0	Lowest threshold: Color–grey-level corresponding to white
OR	
0.10	Color – grey-level corresponding to high-intense yellow
OR	
0.25	Color – grey-level corresponding to low-intense yellow: On transition to green
OR	
0.50	Color – grey-level corresponding to high-intense green
OR	
0.75	Color – grey-level corresponding to low-intense green: On transition to red
OR	
0.90	Color– grey-level corresponding to high-intense red
OR	
1.0	Highest threshold: Color– grey-level corresponding to white

GO TO *Part B*

In order to determine the efficacy of the adoption of Haralick's coefficient (as conceived in this study) in deducing the proportionate extents of yellow, green and red population of spots in a test microarray, three synthetic images constructed from the original microarray of Figure 2 correspond to the following cases: (i) Yellow-rich image; (ii) green-rich image; and (iii) red-rich image. These are illustrated in Figure 7. The choice of images as above stems from the following considerations: As indicated earlier, DNA microarray data in essence depicts differentially genes and gene expression is the term used to describe the transcription of the information included within the DNA, the repository of genetic information, into messenger RNA (mRNA).

In practice, the array chips as indicated before are made available from wet-lab sources with scanning facility etc. In the present study, however only synthetic images are considered to illustrate the underlying analysis. The synthesized images are close hypothetical replicas of laboratory specimens; therefore, the analysis presented can be extended to wet-lab samples without any loss of generality. Further, the synthesized images (Figure 7) are artificially enriched with red, green and yellow spots to stress the algorithmic response to the computations involved.

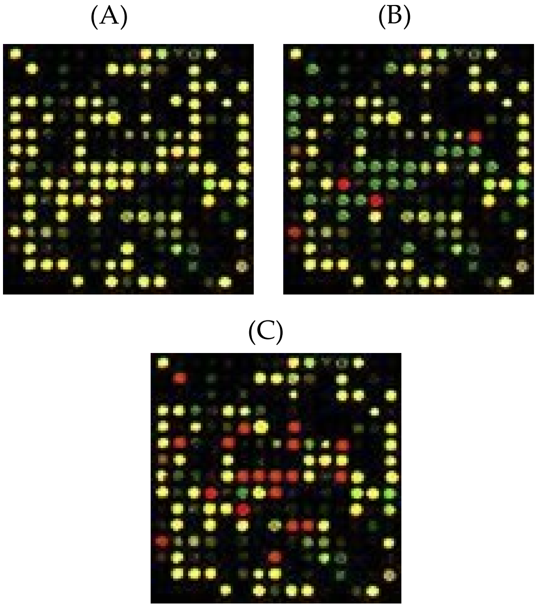


FIG.7 SIMULATED SYNTHETIC MICROARRAYS: (A) YELLOW-RICH IMAGE; (B) GREEN-RICH IMAGE; (C) RED-RICH IMAGE

It can be recalled that all-black pixels correspond to a microarray structure scaled as grey-level '1'; and on the other extreme side, when all the pixels are white, it is graded as '0'. In between, the yellow-rich image would correspond to a grey-level closer to zero, say, 0.1, and the red-rich image would be closer to 1, say 0.9; lastly, the green-rich condition can be taken as a grey-level in the vicinity of 0.5. Results obtained for four different images of the test microarrays illustrated in Figure 7 having varyingthreshold values of grey-scaling in increments of 0.1 and their respective ENT scores are shown in Figure 8.

```
-------------------------------------------------------------------------
Pseudocode # II - Part B:
Computation of GLCM Matrix
Apply
        →   Thresholds to test microarrays
Compute
        →
 %%      Haralick coefficient-based score: Relative extent of grey-
         scale (proportioned to the intensity levels of yellow, green,
         red spots)
        →   Set of probability values, {P(i,j)} each representing
            the implicit value of transition from pixel of an i^th
            grey-intensity to a pixel of a j^th grey-intensity. Scores
            deduced thereof depict normalized relative
            proportion of the colors involved
        →   {P(i,j)} ascertained would be distinct for each color
Read
        →   {P(i,j)} in a file I for each threshold setting, $t_T$ vis-à-vis
            assigned color
Construct
        →   GLCM for each threshold setting, $t_T$
Read
        →   Grey-level values in normalized form expressed in
                terms of {P(i,j} set in file I
        →   Transfer the probability values to a new matrix G
        ←   GLCM for each threshold setting, $t_T$
GO TO           Part C-1
-------------------------------------------------------------------------
Pseudocode # II: Computation of Haralick's parameters - Part C-1
(Loop_1 on GLCM columns )
Define
        →   A compatible Haralick's parameter:
 %       Relevant algorithm is the ENT of equation (1)
Perform
        →   Computation of the Haralick's coefficient:
            Using ENT = I: Eq. (1)
Initialize
        →   Initialize a variable x: i = 0
FOR
        →   Create loop_1 to iterate number of times = Number
            of columns: x = 0, 1,…, (M –1)
SUM
        →   Perform computation
        ←   $I(x) = -\sum_{x} P(x)\log_2[P(x)]$ bits
END
        →   Loop_1
GO TO       Loop_2          ( in Part C-2)
-------------------------------------------------------------------------
Pseudocode # II: Computation of Haralick's parameters - Part C-2
(Loop_1 on GLCM rows )
Initialize
        →   Initialize a variable y: j = 0
FOR
        →   Create loop_2 to iterate number of times equal to
            number of rows: y = 0,…, (N – 1)
SUM
        →   Perform computation
        ←   $I(y) = -\sum_{y} P(y)\log_2[P(y)]$ bits
END
        →   Loop_2
PLOT
            →   Plot S = $\log_e(|$ ◎$|)/I_o$ whereas ◎ I is $(I - I_o)$
                and $I_o$ denotes the average of I
END
-------------------------------------------------------------------------
```

Results and Discussions

It can be seen from Figure 8 that these three sets of microarrays (namely, yellow-rich (b), green-rich (c) and red-rich (d) versions) yield distinct ENT scores *vis-à-vis* the threshold setting chosen. Further, each one of them displays distinct results with respect to that of the original microarray image (taken as such without manipulations on the population of colored pixels). (In Figure 8, logarithm of the differential value of the raw ENT scores is presented in order to observe a good resolution).

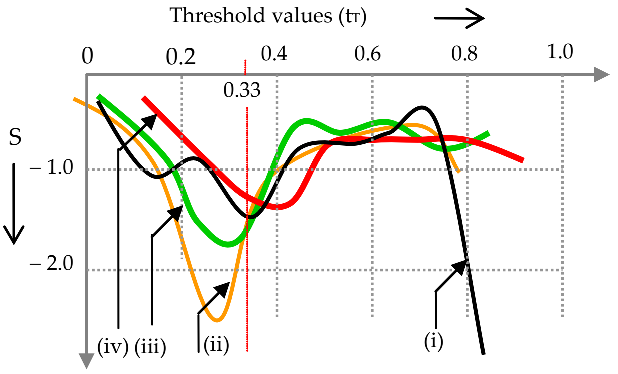


FIG. 8 HARALICK'S COEFFICIENT-BASED SCORE (S) *VERSUS* THRESHOLD VALUE (T_T) INTRODUCED ON TEST MICROARRAYS RESULTS CORRESPOND TO:

FIG. 2 - (i) ORIGINAL RAW IMAGE WITHOUT ALTERATIONS IN COLOR POPULATIONS.

FIG. 7 - (ii) YELLOW-RICH IMAGE; (III) GREEN-RICH IMAGE; AND (IV) RED-RICH IMAGE.

THE ORDINATE SCALE IS TAKEN AS: $S = LOG_E(|◎ I/I_O|)$, WHERE ◎ I IS ($I – I_O$) AND I_O DENOTES THE AVERAGE OF THE ENT (I)

It can be noticed that the score results as seen in Figure 8 are 'peaky' around the threshold $t_T = 0.33$. The reason is as follows: A mixture of three entities namely, the pixels of yellow, green and red colors pertinent to the present study prevails as pixels and coexists in the image-space complying with the heuristics of a stochastical mixture system (Pan, et al., 2003). When the mixture entities are equally populated, the underlying (statistical) partitioning means ideally, a proportion of 1/3. As such, in Figure 8, it can be observed that relevant results on the scores computed, namely, curve (i), exibit a predominant distinguishability around 0.33 (= 1/3) of the threshold value. In the cases of unequal proportions of the mixture constituents, such predominance is seen at t_T value offset (±) from 1/3. For example, curve (ii) is more towards 0 as governed by the rich population of yellow pixels; and curve (iii) is offset more towards 1 as governed by the rich population of red pixels.

Closure

The present study provides a novel approach toward using the Haralick's coefficient to analyze DNA microarrays so as to obtain entropy co-occurrence measures on the relative extent of associated gene expressions. To illustrate the application as above, a typical set of synthetic microarrays is considered, each of which has distinct populations of yellow, green and red fluorescent spots. Such colored features implicitly depict the relative extent gene expression in the array. (For example, green overwhelms the others when the gene expressed corresponds to a non-tumor sample; red significantly prevails when the gene expressed refers to a tumor sample and abundance of yellow signifies when the gene expressed implies both tumor and non-tumor states). Deducing such relative variability of gene expressions has been of primary focus in DNA microarray analyses as evinced by the studies reported, for example in (Newton, et al., 2001; Broberg, 2003; Gill, et al., 2010)

For the analysis proposed and envisaged here, four sets of synthetic images of DNA microarrays are used. The first image is the raw (original) image of the microarray collected from the literature (Nair, 2008; Marquez, et al., 2005). The second image refers to the raw image altered manually with the inclusion of more yellow spots making it yellow-rich. Likewise, the third image is formed to be green-rich and the fourth image is rendered red-rich. For the four test-images as above, corresponding grey-images are constructed with varying thresholds from 0 to 1 in the interval of 0.1. For any given image, a low threshold (→ 0) setting would correspond to the extent of yellow-rich status. Likewise, for a large threshold value setting (→ 1), it is relevant to the status of red-rich condition. For each threshold setting, the GLCM of the image and the Haralick's score (of entropy measures) are obtained. Thus, the plot in Figure 8 of the score value (S) *versus* threshold level illustrates the relative status of a colored pixel population in the microarray being significantly rich or diluted. As mentioned earlier, Cy5 (red) rich profile if observed, it denotes that the sample DNA or cDNA is derived from a tumorous tissue hybridized to the target DNA. Presence of rich Cy3 (green) profile represents the control DNA or cDNA is derived from a normal tissue hybridized to the target DNA; and the yellow color depicts a combination of control and sample DNA both hybridized equally to the target DNA. That is, yellow features have a similar level of expression in both samples.

The present study is indicated adjunct to other possible methods of assaying DNA microarrays for gene expressions in vogue and its efficacy is made evident *via* the distinguishability of color-settings as observed in Figure 8. In addition, the use of Haralick's coefficient in the context of DNA microarray analysis is new and hitherto unexplored. A parallel study akin to the present effort is however, presented in (Neelakanta and Pappusetty, 2012 ; Neelakanta, et al., 2012) where, *in lieu* of Haralick coefficients, a bioinformatics-inspired method is adopted. Yet another new track suggested here as an open-question for research in interpreting DNA microarray patterns is to use an artificial neural network (ANN) with a teacher value prescribed in terms of a Haralick coefficient (Marquez, et al., 2005). Lastly, as regard to the computational cost, it is decided in the present method by the size of the GLCM deduced consistent with the pixel-matrix constructed from the image. For precision and accuracy of classification, a digital image should be formed with larger number of pixels for greater resolution. And corresponding pixel-matrix size will decide the eventual computation burden imposed by GLCM based classifications. Both in the present method as well as in other peers, accuracy of classification versus pixel-matrix size would be the deciding factor on the eventual computational cost.

REFERENCES

Bala Rajni and Agrawal R. K. "Mutual Information and Cross Entropy Framework to Determine Relevant Gene Subset for Cancer Classification", Informatica, 35 (2011), 375-382.

Baggerly K. A., Coombes K. R., Hess K. R., Stivers D. N., Abruzzo L.V., and Zhang W. "Identifying Differentially Expressed Genes in cDNA Microarray Experiments." Journal of Computational Biology, 8(2001), 639-659.

Broberg, Per. "Statistical Methods for Ranking Differentially Expressed Genes." Genome Biology, 4(2003), R41.1-R41.9.

Bryant Penelope, Venter Deon. J., Roy Robins-Browne, M. and Curtis Nigel. "Chips with Everything: DNA Microarrays in Infectious Diseases." The Lancet (Infectious Diseases), 4(2004), 100-111.

Caban Jesus J., Joshi Alark. and Rheingans Penny. "Texture Based Feature Tracking for Effective Time-varying Data Visualization." IEEE Transactions on Visualization and Computer Graphics, 13(2007), 1472-1479.

Chaddad Ahmad, Tanougast Camel., Dandache Abbas. and Bouridane Ahmed. "Extracted Haralick's Texture Features and Morphological Parameters from Segmented Multispectrale Texture Bio-Images for Classification of Colon Cancer Cells." *WSEAS* Transaction on Biology and Biomedicine, 8(2011), 39-50.

Ewiss, A A., et al. "A History of Microarrays in Biomedicine." Expert Review of Molecular Diagnostics, 5(2005), 315-328.

Furlanello Cesare., Serfini Maria, Merler Stefano and Jurman Giuseppe. "Entropy Based Gene Ranking without Selection Bias for the Predictive Classification of Microarray Data." BMC Bioinformatics, 4(2003):1-20

Gill Ryan, Datta Somnath and Datta Susmita. "A Statistical Framework for Differential Network Analysis from Microarray Data." BMC Bioinformatics. 11(2010): 95-(1-10).

Haralick Robert. M., Shanmugam K. and Dinstein Its'hak. "Textural Features for Image Classification." IEEE Transactions on Systems, Man and Cybernetics, SMC-3(1973): 610-621.

Kalpana Ramakrishna and Muttan S. "Assessment of Geriatric-specific Changes in Brain Texture Complexity Using Back Propagation Neural Network Classifier." Complex System, 20(2012): 305-324.

Kalpana Ramakrishna, Muttan S. and Kumarasamy N. "Virus Infection in Brain White Matter: Statistical Analysis of DTMRI Scans." International Journal of Bioinformatics Research and Applications. 7(2011): 227–286.

Marquez Midel C., Perez P. P. and Lagunez-Otero J. "An Evolving Neural Network for the Interpretation of Gene Expression Patterns." Omics: Journal of Integrative Biology, 9(2005): 209-217.

Nair A. J. "Introduction to Biotechnology and Genetic Engineering." New Delhi (India): Infinity Science Press LLC, 2008.

Neelakanta Perambur S. and Pappusetty Deepti. "Bioinformatics Inspired Algorithms for 2D-image Analysis—Application to Synthetic and Medical Images

Part I: Images in Rectangular Format." International Journal of Biomedical and Clinical Engineering, 1(2012): 14-38.

Neelakanta Perambur S., Bertot Edward M. and Pappusetty Deepti, "Bioinformatics Inspired Algorithms for 2D-image Analysis-Application to Medical Images Part II: Images in Circular Format." International Journal of Biomedical and Clinical Engineering, 1(2012): 49-58.

Newton Michael A., Kendziorski Christina .M., Richmond C.S., Blattner F.R. and Tsui, Kam-Wah. "On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data." Journal of Computational Biology, 8(2001), 37-52.

Pan Wei, Lin Jizhen and Le Chap T. "A Mixture Model Approach to Detecting Differentially Expressed Genes with Microarray Data." 2(2003): 117-124.

Schena M. "Microarray Analysis", John Wiley & Sons, Hoboken: New Jersey, 2003.

Shapiro Linda G. and Stockman George C. "Computer Vision." Upper Saddle River: NJ, Prentice-Hall Inc., 2001.

Stekel Dov. "Microarray Bioinformatics", Cambridge UK: Cambridge University Press, 2003.

Wang, Yi and Yan Hong. "Entropy Based Sub-dimension Evaluation and Selection Method for DNA Microarray Data Classification." Bioinformation, 3(2008): 124-129.

Ward Kenneth., "Microarray Technology in Obstetrics and Gynecology: A Guide for Clinicians." American Journal of Obstetrics and Gynecology, 195(2006): 364-372.

Zhu Shenghuo, Wang Dingding, Yu Kai, Li Tao and Gong Yihong. "Feature Selection for Gene Expression Using Model-based Entropy." IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7(2010):25-36.

Kalpana Ramakrishnan Ph. D (2012) degree (in electrical engineering) from Anna University, Chennai (India). She is currently Professor at Rajalakshmi Engineering College, Chennai, India. Her area of research interest includes bioinformatics, image processing and nano sensors.

Dr. R. Kalpana is a Member of ACEEE and IEI, Calcutta.

Perambur S. Neelakanta received his Ph. D. degree (in electrical engineering) from Indian Institute of Technology, Madras (Chennai), India in 1975.

He is currently Professor in the Department of Computer & Electrical Engineering and Computer Science (CEECS), Florida Atlantic University (FAU), Boca Raton, Florida 33431, USA. His areas of specializations include: Bioinformatics, neural networks, Mathematical biology, bioelectromagnetics.

Dr. Neelakanta is a Chartered Engineer (UK) and a Fellow of IEE (UK), (now known as IET).